Proceedings of the Asia-Pacific Advanced Network 2013 v. 36, p. 98-105
http://dx.doi.org/10.7125/APAN.36.14
ISSN 2227-3026

Collecting and Grouping of Distributed and Heterogeneous SNS Contents for Collaborative Storytelling

Abbas Ali Butt [1], Jaehyuk Park [2] and Yong-Moo Kwon [3,*]

1 KIST / 39-1 Hawalgogdong Sungbukku, Seoul, KOREA
2 KIST / 39-1 Hawalgogdong Sungbukku, Seoul, KOREA
3 KIST / 39-1 Hawalgogdong Sungbukku, Seoul, KOREA

E-Mails: abbas395@hotmail.com; leopark83@gmail.com; ymk@kist.re.kr

* Tel.: +82-2-958-5767; Fax: +82-2-958-5769

Abstract: In this paper we describe how to collect and group the contents of distributed and heterogeneous SNSs for storytelling. In particular, we have addressed two issues for storytelling, i.e. collecting and grouping. At first, when users input their queries, our proposed system collects proper contents from distributed and heterogeneous SNSs using content-based filtering, followed by its grouping, based on similarity between content text information and user query. Our implementation results show that a more informative and detailed online story can be made by using SNS contents from Facebook and Twitter. Moreover, these contents can be grouped automatically rather than manually.

Keywords: Storytelling; Social Curation; Distributed and Heterogeneous SNS; Social Factor; SNS Feature.

1. Introduction

In the current era, social network services (SNSs) have become important media to stay in touch and to share experiences and activities of users' daily lives. The statistics presented by Hachman claim that Facebook has 901 million users [1]. Parr reported that 250 million photos are uploaded every day on Facebook [2]. On Twitter, people post around 750 tweets per second [3]. Through these SNSs, users can directly share photos, opinion and information. Moreover, users can also add more information to their uploaded photos to make it more meaningful such as, making albums, providing captions, adding tags about related people, and leaving comments on other's contents. Not only the photo author can add information, his SNSs' friends can give user

responses, including the form of likes, comments and tags. All these operations are for semantic enrichment of these social contents and for making a story of our lives as valuable sources

Some SNS contents creation studies have been reported by researchers regarding recommendation and storytelling techniques [4-5]. Specifically [4] provides the content recommendation by using Collaborative Authoring Metadata (CAM). And in [5], authors collected and grouped the single SNS social content by using the technique Expectation-Maximization algorithm. In this work, we collected distributed and heterogeneous SNS contents and grouped them on the bases of user query and content similarity with semantic information.

2. Approach

Figure 1 shows the overview of our system. Our application will target those events which can have some sub events like picnic, hiking or sch. excursion

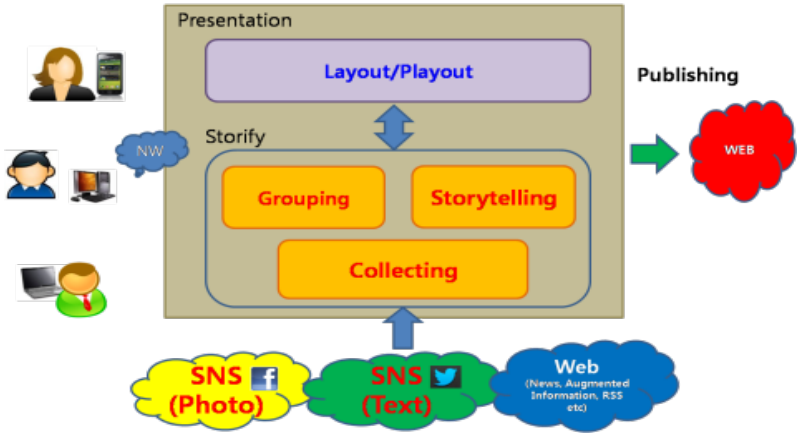

Figure 1. Overview of our social curation system for storytelling.

So for this purpose, we will use the 5W1H technique. We use 5W1H as follows; "When" - means date (since and until) of the event, "Who" - means the participants of the event, "What" - means the title of the event, "Where" - means the location of the event, "Why" - means the purpose of the event and "How" means such as sub events of the main event. For example, users went to the Korean Cultural Experience Tour 2013 at 7th April in Seoul with his college friends and experience the Blue House, Korean traditional tea, dress and music. In this example 7th April (since = 7th April, until = 7th April) means one day tour, college friends like Mahmoud, Fattahi, Maria etc, Korean Cultural Experience Tour 2013, Seoul, experience the Korean Culture for When, Who, What, Where and Why respectively . And Blue House, Korean traditional tea, dress and music will be the "How" of the event.

Facebook image features such as likes, comments, participants (unique persons who react), tagged persons, caption, and timestamp and Twitter tweet features such as text, favorite count, re-tweet count and comments features can be used for the collecting purpose.

On the base of these Facebook and Twitter features, we calculate the DIW (Degree of Interests Width) and DPD (Degree of Participation Depth). As we know, the comments and likes show how much people are interested in the content. So we can identify the content DIW and DPD by using the number of comments and likes of the content. Table 1 shows how to calculate the DIW and DPD values based on metadata in Facebook and Twitter.

Table 1. Calculation of DIW and DPD for both Facebook and Twitter.

Media	Name	Calculation
Twitter	DIW	Num. of mentions from non-participants + Num. of "Favorite"s + Num. of "RT"s
	DPD	Num. of mentions among participants + Num. of participants
Facebook	DIW	(0.5 * Num. of "Like"s) + Num. of commenters + (2 * Num. of "Share"s)
	DPD	(Num. of total comments / Num. of commenters) + Num. of commenters

2.1. Content Collecting from distributed and heterogeneous SNSs

In this section we will describe how semantic information in social contents can be interpreted from a user query. In case of Facebook, a user query includes target event dates, location and participants. In case of Twitter the query can include date and user's id. We map user query 5W1H to the semantic information of social contents as follows.

- Date of the content should be between the since and until date of the user query
- Location of the event should belong to the location defined by the user. E.g. user said Korea then content that have location Seoul or Busan belongs to the same event.
- Album title or image caption should be similar to the user query parameter of what, why, how.
- Tagged persons of the content can be mapped to the parameter of the user query. Basically, Facebook photos of a user can belong to one of four categories: Profile photos which are a display picture of the user, Cover photos which show on the large area of user timeline [6], Wall photos in which user or his friends share on the wall, and Album photos in which user takes on different events and so uploads at Facebook on their behalf in the form of albums. Among those categories, Album photos owned by a user and his friends contain valuable semantic information. Therefore we mainly focus on Album photos of a user and his friends for our storytelling.
- For the text content we use Twitter[1]. Twitter messages, which are text contents in the message, the number of marked as a favorite and re-tweeted of the message.

[1] So we assume user will make the stories related to the recent events so tweet content related to the event is included in his recent 3200 tweets that we can fetch through Twitter API. Specifically reference [7] author says user is more interested to make the story of recent events so our assumption is reasonable.

Our general approach for content collecting is based on 5W1H (When, Who, What, Why, How). Algorithm 1 and 2 shows how to collect the SNS contents from Facebook and Twitter respectively.

```
all_albums ← getFBContents(since,untill)
current_album_no ← 1
related_albums ← {}
while(current_album_no < sizeOf(all_albums))
        album ← all_albums[current_album_no]
        participants ← getParticipants(album)
        if((particpants ∩ person_query) ≠ ⦰ OR isSimilar(getAlbumTitle(album),what) OR
isSimilar(getAlbumTitle(album),why) OR isSimilar(getAlbumTitle(album),How))
                related_albums = related_albums ∪ album
        current_album_no++
```

Algorithm 1. The algorithm of selecting the related contents from Facebook Here, getParticipants(album) will return all those participant names that are tagged in the album or any image of the album. It should be noted that one person can present at one location at a time. Therefore, the tagged person information is useful to collect the related photos from Facebook. If the person is tagged in the photo, it indicates he/she participate in the event on that date.

```
all_tweets ← getTwitterTweets(diplayname)
current_tweet_no ← 1
related_tweets ← {}
while(current_tweet_no < sizeOf(all_tweets))
        tweet ← all_albums[current_album_no]
        if((tweet['text] ∩ what) ≠ ⦰ OR ((tweet['text] ∩ why)) ≠ ⦰ OR ((tweet['text] ∩ how)) ≠ ⦰ OR
isSimilar(tweet['text] , where))
                related_tweets = related_tweets ∪ tweet
        current_tweet_no ++
```

Algorithm 2. The algorithm of selecting the related contents from Twitter Here, getTwitterTweets (displayname) will return recent 3200 tweets of the user. Then we will filter the tweets according to the date. In next step we will check the similarity between tweet text and other common query parameters. If tweet text similar to one of these parameters then we will select that tweet.

After collecting SNS contents, SNS user feedback information are used for selecting or ranking them. Facebook's user feedback includes the number of "Like"s, the number of comments and tagged people, and the number of participants. Twitter's user feedback includes favorite count and re-tweet count. We classify these feedbacks into two parameter, i.e., DIW (Degree of Interests Width) and DPD (Degree of Participation Depth). The DIW includes the number of likes, the number of commenters, and the number of shares in Facebook while, for Twitter, it includes the Favorite and Re-tweet counts and the number of mentions (the casual format of Twitter messages) from non-participants. On the other hand, the DPD includes the average number of comments per commenter and the number of commenter in Facebook, while the number of mentions among participants and the number of participants in Twitter.

After measuring DIW and DPD for each content, we collect additional data using Google API according to the following process. First, we selected the five contents from both Twitter and Facebook which have highest DIW values and DPD values, respectively. Second, for the selected between five and ten photo contents from Facebook, similar images are searched through Google Image API with titles of the selected photos as keywords. At the same time, for the selected between five and ten tweet contents from Twitter, related news articles are searched through Google News API with texts in the tweets. Figure 2 shows our whole collecting process from Facebook, Twitter and Google.

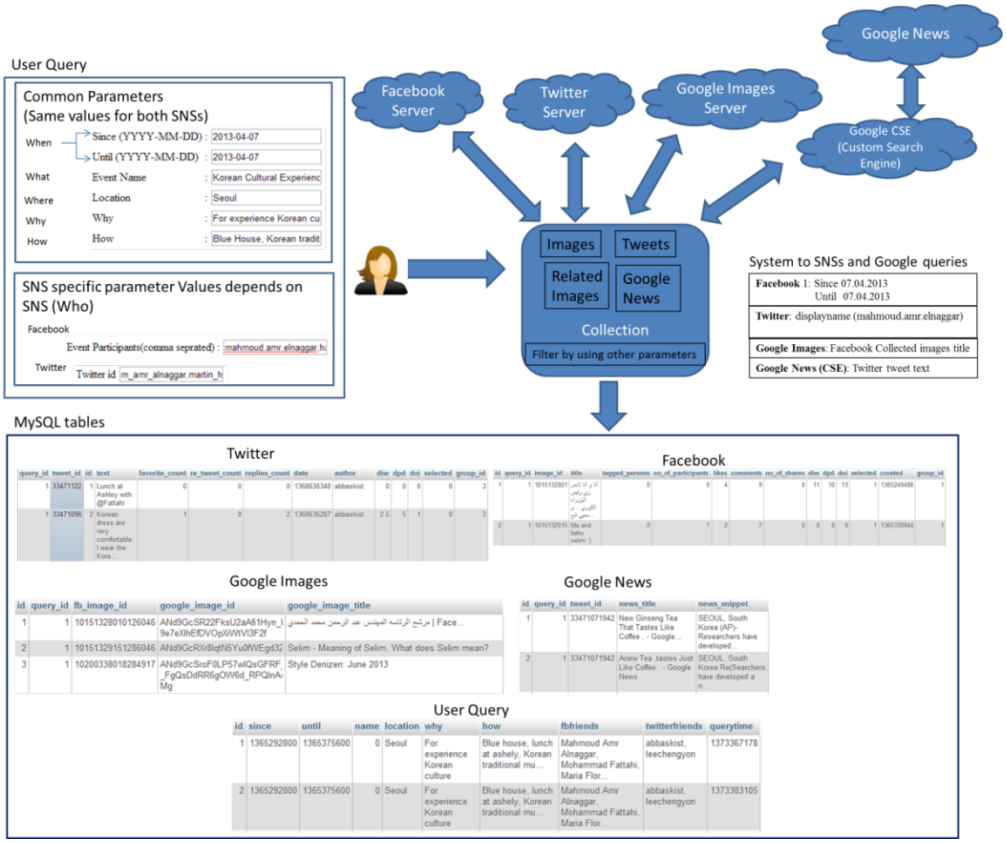

Figure 2. Content Collecting Process from Facebook, Twitter and Google.

2.2 Content Grouping

In this section we will describe how we group the selected images and tweets according to the user query parameter "How". Here, "How" can indicate the sub-event names. So, we can group the SNS contents into several sub-event groups. It should be noted that the sub-event contents can be grouped by using sub-event title and sub-event time duration information. This sub-event information can be used at the plot process in storytelling.

We use the k-means clustering method for grouping the collected contents. Before applying k-means, we normalize the DIW, DPD values and posted time points of collected contents. Then, as the first step of clustering, we initialize the center of each dimension (DIW, DPD, time and text). For the centers of the text dimension, according to the 'How' parameter of a user query, we set the same number of centers and assign texts in the 'How' parameter to each centers as names.

For each center of other three dimension values, we subtract minimum value from the maximum value of each dimension then divide it by the number of centers. Then we assign this value to the first center, double value of the first center as the second center value and so on. For example, assume that there are five centers of one dimension having the minimum value 0 and the maximum value 1. Then for the first center value we assign will be (1-0)/5 = 0.2, and 0.4, 0.6, 0.8 and 1.0 for 2nd, 3rd, 4th and 5th respectively.

After assigning center values for our four dimensions, we clustered those using k-means clustering method.

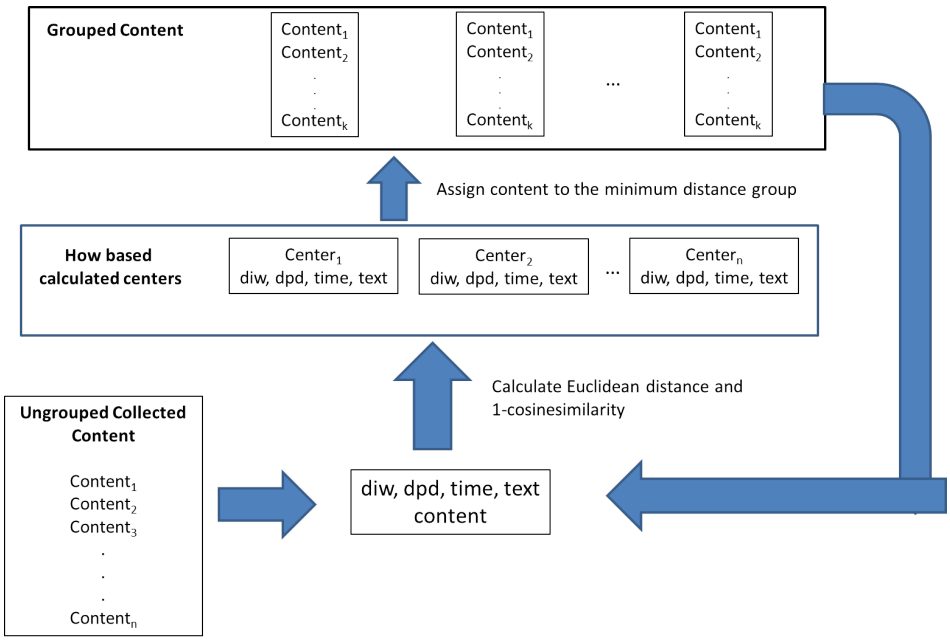

Figure 3. Concept of grouping of the collected content.

3. Results

Our implementation is still under process. Figure 4 shows how the images and text media from Facebook and twitter are collected and the result of the grouped content.

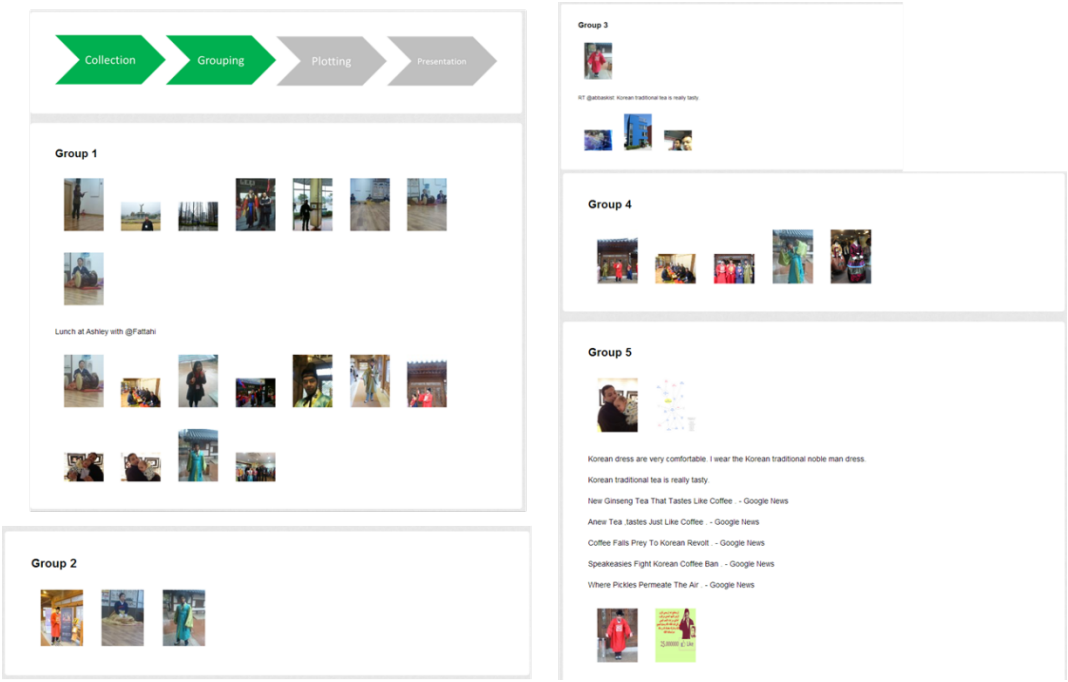

Figure 4. Step-1 Content Collecting result.

4. Conclusions

Currently, we are now developing storytelling system using social curation technique. This paper describes how to collect and group the SNS contents from distributed and heterogeneous SNS contents. In the collecting stage, the concept of 5W1H and social factor of SNS are used. Then by using k-means clustering method based on the user query parameter “How”, the collected SNS contents are clustered into sub-event groups. The next step includes how to plot the grouping SNS contents in storytelling, while utilizing the DIW and DPD values as well as SNS features that are tagged in the photo and the participated person name in the comment and tweet.

Acknowledgements

This work is supported by “Development of Tangible Social Media Platform Technology” of KIST.

References

1. Hachman, M., (posted 23.04.2012). Facebook Now Totals 901 Million Users, Profits Slip. Web Page. http://www.pcmag.com/article2/0,2817,2403410,00.asp
2. Parr, B. (posted 21.10.2011), Infographic: Facebook by The Numbers, Web Page, http://mashable.com/2011/10/21/facebook-infographic/
3. Go-Globe.com (posted 30.10.2012), Social Media Statistics and Facts 2012, Web Page, http://www.go-globe.com/blog/social-media-facts/
4. Fathoni Arief Musyaffa (2012), SNS-based Content Recommendation Scheme for Web-based Social Collaborative Authoring(MS Thesis), University of Science & Technology
5. Mohamad Rabbath, Philipp Sandhaus, and Susanne Boll (2010), Automatic creation of photo books from stories in social media. In *Proceedings of second ACM SIGMM workshop on Social media*(WSM '10), 15-20
6. Facebook(2013), How do I add or change my cover photo?, Web Page, https://www.facebook.com/help/220070894714080
7. Nancy A. Van House (2009). Collocated photo sharing, story-telling, and the performance of self. Int. J. Human-Computer Studies 67, 1073–1086.